European Scientific Journal July 2015 edition vol.11, No.20 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431

POWER LAW DISTRIBUTION AS A COMPONENT OF THE VERTEX DEGREE DISTRIBUTION ON A SOCIAL UNIVERSITY NETWORK COURSE

Orgeta Gjermëni*, *PhD Candidate
University "Ismail Qemali", Vlore, Albania

Abstract

The aim of this paper is to analyze a collection of data gathered from surveys held every three weeks in a Spring Course of the Economic Faculty in the University "Ismail Qemali"of Vlora, Albania. The data set for each student also contains the names of other students through which he/she have a "social relationship". This social relationship includes frequent communications, discussions on exercise solutions, and sitting usually close to each other in the class. We have constructed four social simple graphs and have analyzed them focusing only on degrees. In addition, we fit discrete power law degree distribution on the tail and their evolution through time. In analyzing the data, we employed the R platform.

Keywords: Social Network, Vertex Degree, Power Law

Introduction

Social network is a field which has emerged from social psychology, sociology, statistics, and graph theory since 1930's. Jacob Moreno was the first who studied the interpersonal relationships. Mathematical formalization happened in the 1950's while theory and methods of social networks became pervasive in the social and behavioral sciences during the 1980's (Wasserman & Faust, 1994; Freeman, 2004). Thus, a social network is a set of people or groups of people with some pattern of contacts or interaction between them (Scott, 2000; Wasserman & Faust, 1994).

However, modern social network analysis is a computationally intensive affair nowadays. It is also related to big data. "Computational Social Science" (Lazer *et al.*, 2009; Kolaczyk, 2009; Kolaczyk & Csárdi, 2014) is a new discipline, which has emerged to join the efforts of social scientists, computer scientist, mathematicians, and physicists in an interdisciplinary approach, with the purpose of better understanding the

behaviors laws of human society at both the individual and collective level. The dependence of network analysis on computations for research has helped in spawning a wide array of software packages for performing analytic tasks such as: R, Graphiz, Pajek, Cytoscape, and Gephi. In our work, we have chosen the open statistical computing platform, R (R Core Team, 2015).

Social networks as real – world random graphs, which evolve in time, are studied mostly on static snapshots at various points in time. Subsequently, these snapshots are used to make inferences about the evolutionary processes. Hence, various studies (Newman, 2003; Guimerà *et al*., 2006; Kumar *et al*., 2006; Onnela *et al*., 2007) were carried out in this field.

Methods

Data was obtained from four surveys which were held every three weeks in a Spring Statistic Course for the second year students of the Economic Faculty in the University "Ismail Qemali" of Vlora. At the beginning of each course, students have the possibility to choose between some alternatives of the lecturer they want. A "mixing process" usually happens at the beginning of every course within the various groups of students. The first survey was held after the course began after three weeks. During the survey, each of the students provided the names of other students which he/she has a "social relationship" with. This social relationship includes frequent communications, discussions on exercise solutions, and sitting usually close to each other in the class. Therefore, this "relationship" defines the socialization that happens within the university course. The surveys which were conducted were considered as "snapshots" at four different moments during the time of the survey. If a "social relationship" starts between two students, it is considered to be "forever" until the course ends.

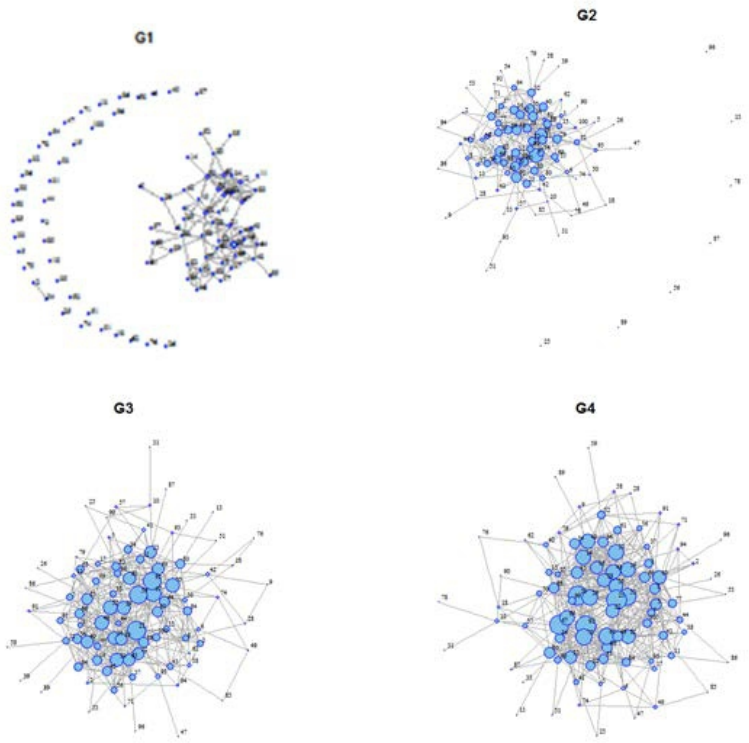

Figure1: Visualization of evolving process in the social graph captured at four moments.

Furthermore, we will apply the mathematical definition of the network, "graph", since analysis is focused on the mathematical context. Our social graph is conceived as a fixed set of vertices. In addition, it is seen at four moments $G_i = (V_i, E_i)$, where $i = \overline{1,4}$ (see Figure 1). Thus, for every $1 \leq i < 4$, $G_i \subseteq G_{i+1}$. In this way, we have the possibility to study the evolving process that happens between the relations held in this course. Let $V_i -$ be the set of vertices (every vertex identifies a student) which is fixed ($|V_i| =$89) from the beginning of the semester, and $E_i -$ be the set of edges (every edge identifies a relationship between two students) which has the tendency to grow pass the time ($|E_1| = 128, |E_2| = 289, |E_3| = 414, |E_4| = 497$). Hence, the orientation of the edges is neglected and the graphs are simplified. *Degree* of a vertex can be defined as the number of incident edges with it. The connectivity of a graph is related with the degree distribution (*probability mass function* (pmf) in the discrete distribution or *probability density function* (pdf) in the continuous distribution), $p(x) = P(X_i = x)$, which measures the probability that a random vertex will have degree equal to x. The *cumulative distribution function* (cdf), $F_i(x) = P(X_i \leq x)$ measures the probability of degrees less than x. The

complementary cumulative distribution function (ccdf), $\overline{\mathrm{F_1(x)}} = \mathrm{P(X_i > x)}$, measures the probability of degrees greater than x .

During the study, we investigated the presence of a power law component on the "tail" of the degree distribution (for x greater than a minimal value x_{min}), given that CCDF plotted on logarithmic scale on both axes has a linear decrease component on the "tail" (see Figure 3). However, we cannot be sure of the presence of a power law known as "linear component". Therefore, what we see in Figure 3 is a necessary condition but not a sufficient one!

Vertex degree is a discrete random variable. As a result, we focus on the investigation of discrete power law. The probability mass function (pmf) of the discrete power law is expressed by the formula below:

$$p(x) = C\mathrm{x}^{-\alpha}, \qquad (1)$$

where ku $\mathrm{C} = \frac{1}{\zeta(\alpha,\mathrm{x_{min}})}$, is the normalization constant in way that $\sum_{\mathrm{x=x_{min}}}^{\infty} C\ p(x) = 1$.

$$\zeta(\alpha, \mathrm{x_{min}}) = \sum_{\mathrm{n=0}}^{\infty}(\mathrm{n} + \mathrm{x_{min}})^{-\alpha}, \qquad (2)$$

Consequently, the above equation is the generalized function or Hurwitz zeta function (Abramowitz & Stegun, 1972). The complementary cumulative distribution function (ccdf) is given as:

$$\overline{F(x)} = \frac{\zeta(\alpha,\mathrm{x})}{\zeta(\alpha,\mathrm{x_{min}})}. \qquad (3)$$

Estimations on discrete parametric power law which better fits on the empirical data, were done as explained by Clauset *et al*. (2009). Furthermore, it was implemented on poweRlaw package (Gillespie, 2014; Gillespie, 2015). The $\hat{x}_{min}$ value estimated is chosen in way that the estimated power law model gets a best fit of the empirical probability distribution for $x \geq x_{min}$ (Clauset *et al*., 2007). Under the supposition that our data follows a power law for $x \geq x_{min}$,, the α parameter is estimated by a numeric optimization of the log – likelihood. In the discrete power law, $\hat{\alpha}$ is approximated with$\hat{\alpha} \cong 1 + \mathrm{n}\left[\sum_{i=1}^{n} ln\frac{x_i}{x_{min}-1/2}\right]^{-1}$. Moreover, the details of this approximation are given by Clauset *et al*. (2009). To estimate the distance between the two model distributions, the empirical and theoretical power law uses the Kolmogorov – Smirnov statistic (KS) (Press *et al.*, 1992).

$$\boldsymbol{D} = max_{x \geq x_{min}}\left|S(x) - \overline{F(x)}\right|, , \qquad (4)$$

where S(x) is the empirical CCDF, while $\overline{F(x)}$ is the theoretical CCDF of the power law model which best fits the empirical data for $x \geq x_{min}$. Estimation for $\hat{x}_{min}$ is defined by the value which minimizes D. Uncertainty on x_{min} estimation is investigated by the non parametric bootstrap method (Efron & Tibshirani, 1993) with 5000 iterations. To estimate α and x_{min} during bootstraps analyses, the maximum likelihood estimation (MLE) is used.

For every data set, it is possible to make a power law fit, despite the fact that the real data are not following this distribution. For this reason, we need to make a test between two hypotheses:

H_0: data is generated according to a power law distribution for $x \geq x_{min}$.

H_1: data is not generated according to a power law distribution for $x \geq x_{min}$.

Hypothesis testing procedure is done as explained by Clauset et al. (2009) based on a *goodness – of – fit test.* This test however generates a p – value that quantifies the plausibility of the hypothesis. Such tests are based on measurements of the "distance" between the distribution of empirical data and the hypothesized model. Consequently, H_0 hypothesis is ruled out for $p \leq 0.1$.

The statistical packages used for analysis include: *igraph* (Csárdi & Nepusz, 2006), *poweRlaw* (Gillespie, 2014; Gillespie, 2015), and *network* (Butts, 2008; Butts, 2015) on the R platform.

Results

To have a general overview on degree evolving process, the evolving degree diagram (see Figure 2) was constructed. Passing from G1 to G2, there is an increase by 3.629 on average to the degrees of vertices from G2 to G3 with 2.798. Furthermore, it also increases from G3 to G4 with 1.865. It is noted that over time, there is an average growth decline on degrees per vertex. After plotting all the complementary cumulative distribution functions (CCDF) on the logarithmic scale on both axes (see Figure 3), it was noted that a "linear" decline which suggests raising the hypotheses was present.

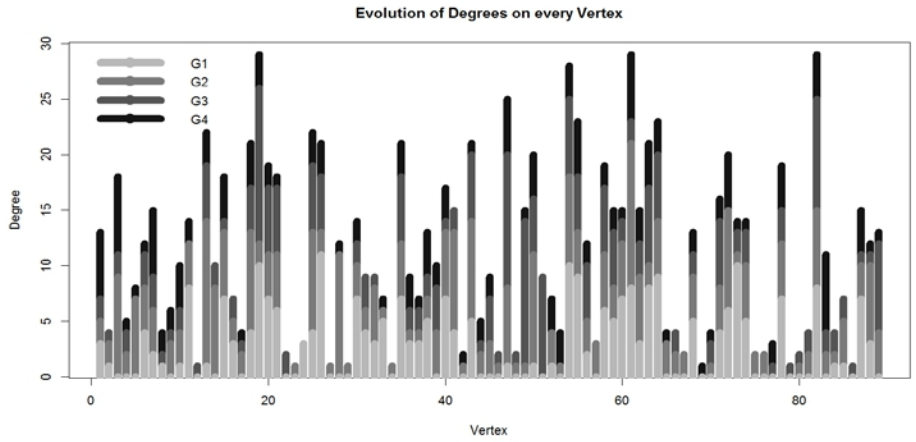

Figure 2: Evolving degree diagram for each vertex of the social graph at four moments.

In addition, a power law component is present especially on the "tail" of the degree distribution. For each of the graphs (G1, G2, G3 and G4), an

estimation of best fitted power law is done. Also, the bootstrap procedure with 5000 iterations is also applied. The results of this procedure are given in Table 1. Also, we have visualized the CCDF's histogram of x_{min}, histogram of α, and the scatter plot of x_{min} versus α, based on the bootstrap procedure with 5000 iterations (see Figure 4, 5, 6, and 7).

Table 1: Results gained through the fitting procedure of discrete power law model to the empirical degree distribution based also on the bootstrap procedure with 5000 iterations.

Distribution	Graph	KS	x_{min}	Sd(x_{min})	α	Sd(α)	p - value
Power law (Discrete)	G_1	0.0639	7	1.187	5.85	1.954	0.58
	G_2	0.0661	13	1.41	8.66	2.27	0.81
	G_3	0.0846	17	2.72	7.05	2.09	0.59
	G_4	0.1282	20	3.648	6.85	2.2	0.14

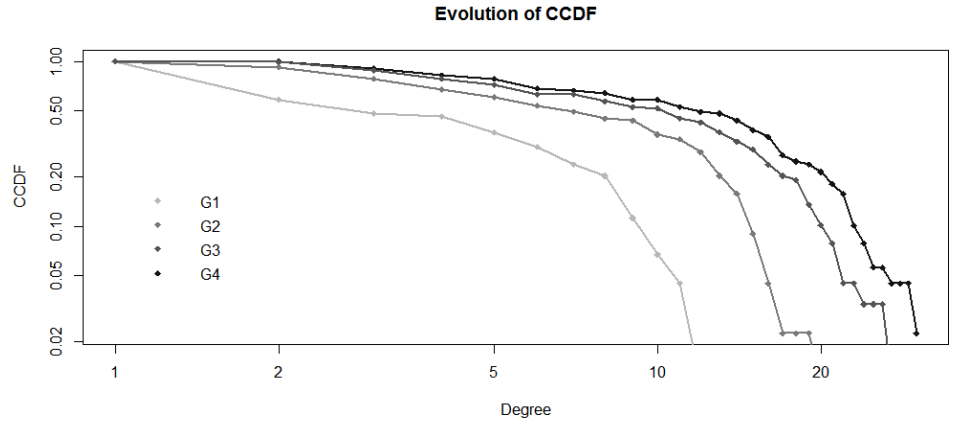

Figure 3: Complementary cumulative distribution functions (CCDF) given at four moments, plotted on logarithmic scale on both axes. Zero degrees are not considered.

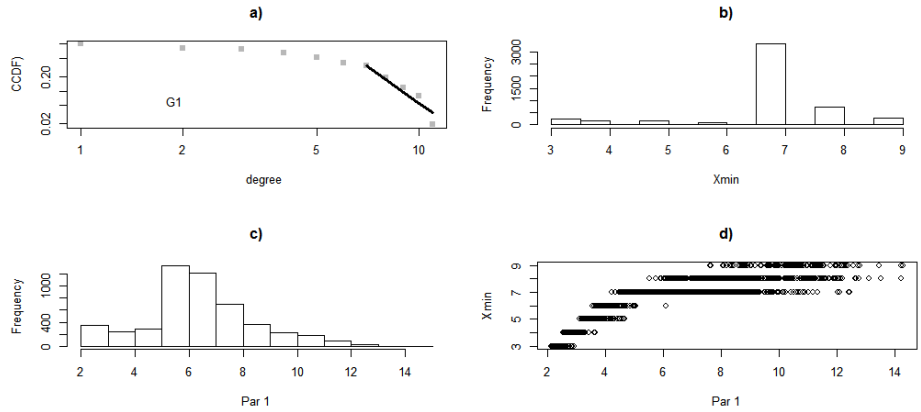

Figure 4: a) CCDF of the data set and the fitted power law linear line of G1 with $x_{min} = 7$ and parameter $\alpha = 5.85$; b) histogram of x_{min} (sd= 1.187); c) histogram of the distribution parameter (Par 1) = α (sd=1.954); and d) scatter plot of parameter α versusx_{min}.

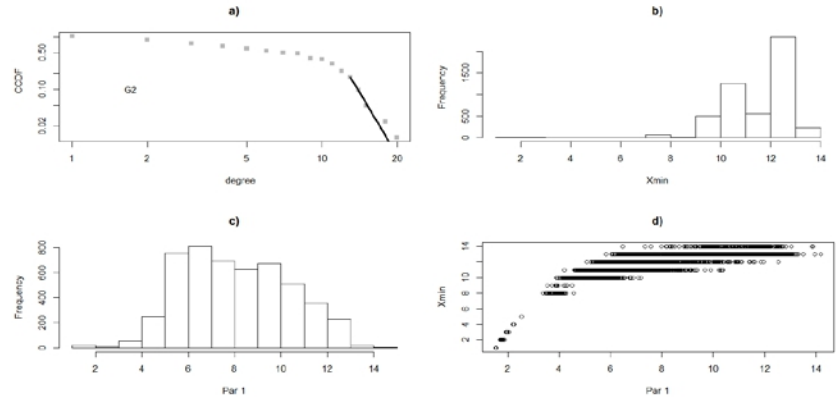

Figure 5: a) CCDF of the data set and the fitted power law linear line of G2 with $x_{min} = 13$ and parameter $\alpha = 8.66$; b) histogram of x_{min} (sd= 1.41); c) histogram of the distribution parameter (Par 1) = α (sd=2.27); and d) scatter plot of parameter α versus x_{min}.

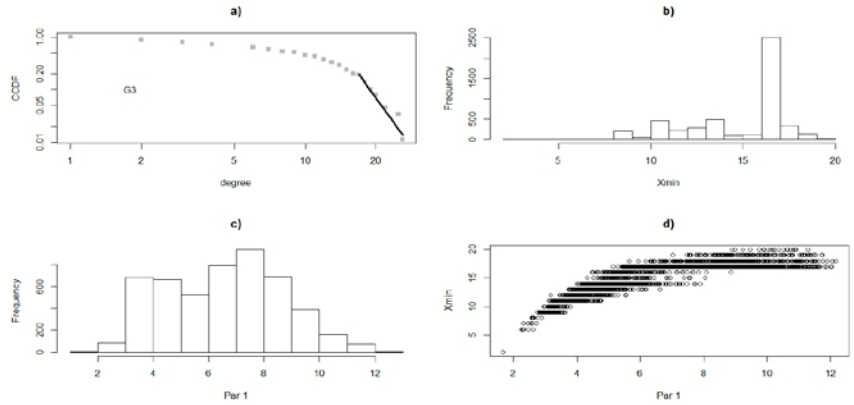

Figure 6: a) CCDF of the data set and the fitted power law linear line of G3 with $x_{min} = 17$ and parameter $\alpha = 7.05$; b) histogram of x_{min} (sd= 2.72); c) histogram of the distribution parameter (Par 1) = α (sd=2.09); and d) scatter plot of parameter α versus x_{min}.

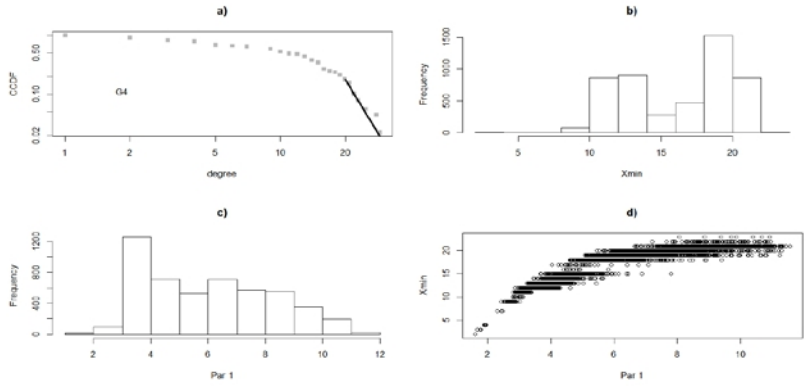

Figure 7: a) CCDF of the data set and the fitted power law linear line of G4 with $x_{min} = 20$ and parameter $\alpha = 6.85$; b) histogram of x_{min} (sd= 3.648); c) histogram of the distribution parameter (Par 1) = α (sd=2.2); and d) scatter plot of parameter α versus x_{min}.

Conclusion and Discussion

During this study, the student's ability to socialize was observed. Thus, socialization among students was greater at the beginning of the course than at the end. Hypothesis raised about the presence of a power law component in degree distributions of the social graph at all the four moments, are not ruled out according to the p – values (see Table 1). Although in the case of G4, p – value (0.14) is too close to the critical value which is 0.1. For the minimal degree values x_{min} for which it was estimated, a theoretical discrete power law distribution for the empirical data, in each of the cases have a tendency to grow passing the time. However, the estimated values of the parameter α fluctuate with time. Since the size of the data (89) is less than 100, discrete power law distribution for $x \geq x_{min}$, ought to be compared with other distributions which can also obtain good fits on the "tail" of the empirical distribution. Therefore, the challenge is to model all the data, and not only the "tail" regardless of the complexity in the evolving process and the network.

References:

Csárdi, G., & Nepusz, T. (2006). The igraph software package for complex network research, *InterJournal*, Complex Systems 1695. http://igraph.org/

Butts, C. (2015). Network: Classes for Relational Data. *The Statnet Project (<URL:http://www.statnet.org >).* R package version 1.12.0, <URL: CRAN.R-project.org/package=network >

Butts, C. (2008). Network: A Package for Managing Relational Data in R. *Journal of Statistical Software*,24(2). <URL:http://www.jstatsoft.org/v24/i02/paper>

Gillespie, C. S. (2015). Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software*,64(2),1-16. <URL: http://www.jstatsoft.org/v64/i02/>

Gillespie, C. S. (2014). A complete data framework for fitting power law distributions *arXiv*:1408.1554v1 [stat.CO]

R Core Team (2015). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/

Wasserman, S. & Faust, K. (1994). Social Network Analysis in the Social and Behavioral Sciences. *Social Network Analysis: Methods and Applications* (pp. 1–27). Cambridge University Press.

Freeman, L., (2004). *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press.

Newman, M. E. J. (2003). The structure and function of complex networks. *arXiv*:cond-mat/0303516v1[cond-mat.stat-mech]

Guimerà, R., Danon, L., Díaz – Guilera, A., Giralt, F., & Arenas, A. (2006). The real communication network behind the formal chart: Community structure in organizations. *Journal of Economic Behavior & Organization* 61, 653 – 667.

Kumar, R., Novak, J., & Tomkins, A. (2006). Structure and evolution of Online Social Networks. *KDD*'06

Onnela, J.P., Saramaäki, J., Hyvöven, J., Szabó, G., Argollo de Menezes, M., Kaski, K., & Barabási, A.L. (2007). Structure and Tie Strength in Mobile Communication Networks. *arXiv*:physics/0610104 [physics.soc-ph]

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. L., Brewer, D. Alstyne, M. V. (2009). Computational Social Science. *Science* 323:721-723

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. New York, NY: Springer. doi:10.1007/978-0-387-88146-1

Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical Analysis of Network Data with R.* Springer – Verlag New York. doi: 10.1007/978-1-4939-0983-4

Scott, J. (2nd Ed.). (2000). *Social Network Analysis: A Handbook*. Sage publications, London.

Abramowitz, M., & Stegun, I. A. (10th. Ed.). (1972). *Handbook of Mathematical Function with Formulas, Graphs, and Mathematical Tables*. Vol. 55 of Applied Mathematics Series.

Clauset, A., Young, A., & Gleditsch, K. S. (2007). On the Frequency of Severy Terrorist Events. *Journal of Conflict Resolution.* 51(1): 58 – 57

Clauset, A., Shalizi, C.R., & Newman, M.E.J., (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703

Press, W. H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (2nd Ed.) (1992). *Numerical Recipes in C: Art of Scientific Computing*. Cambridge University Press, Cambridge, England

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.